

Environment-Aware Predictive Sampling for Base Station Selection in Non-Stationary A2G Channels

Jiebin Xie, Rui Zhang, Suzhi Bi and Zhi Quan

College of Electronics and Information Engineering, Shenzhen University, China

Email: xiejiebin2023@email.szu.edu.cn, {zhrrygh, bsz, zquan}@szu.edu.cn

Abstract—Unmanned aerial vehicles (UAVs) need to switch among ground base stations (GBSs) under fast three-dimensional mobility, which drives strong non-stationarity in air-to-ground (A2G) links. Classical multi-armed bandits often incur large dynamic regret, whereas deep models are costly on payload-limited platforms. This paper proposes environment-aware predictive sampling (EA-PS), which adopts a hierarchical model that multiplies a radio-map large-scale prior, a spatial Gaussian process (GP) residual for shadowing mismatch, and a small-scale mean governed by first-order auto-regressive (AR(1)) dynamics within the predictive sampling (PS) framework. Independent posterior samples of the residual and the small-scale state are merged into the arm-selection rule, preserving the reversible PS structure while enabling joint spatiotemporal exploration. We derive a Bayesian regret upper bound through a mutual-information decomposition that separates GP-induced spatial uncertainty from AR(1) temporal uncertainty. QuaDRiGa simulations under the 3GPP Urban Macro scenario show about 23% higher cumulative reward and over 50% lower cumulative regret relative to representative non-stationary bandits.

Index Terms—Air-to-ground channel, Gaussian process, predictive sampling, non-stationary bandits, ground base station selection.

I. INTRODUCTION

In the vision of sixth-generation (6G) communications, realizing the Internet of Everything and space-air-ground integrated networks has emerged as a fundamental objective. Owing to their flexible deployment and high likelihood of establishing line-of-sight links, unmanned aerial vehicles (UAVs) are indispensable in various 6G applications, such as intelligent transportation systems and large-scale data collection [1]. However, rapid three-dimensional maneuvers require UAVs to switch among multiple ground base stations (GBSs) to maintain high link quality. Meanwhile, fast topology changes and inherent mechanical jitter introduce strong non-stationarity into air-to-ground (A2G) channels [2], [3]. These effects exacerbate online GBS selection. Therefore, to accommodate stringent payload and power constraints, it is imperative to develop an efficient, low-complexity online algorithm.

Classical online learning frameworks, particularly multi-armed bandit (MAB) models, are widely employed to balance exploration and exploitation in base station selection [4].

However, traditional MAB algorithms struggle with rapid topological variation in A2G channels and often incur linear dynamic regret. While deep reinforcement learning [5] and deep recurrent sequence models [6] can track complex temporal dynamics, their computational overhead, low sample efficiency, and susceptibility to catastrophic forgetting render them impractical for resource-constrained UAVs. As a lightweight alternative, predictive sampling (PS) [7] has emerged as an effective algorithm for non-stationary bandits. However, PS inherently treats non-stationarity as purely temporal evolution (for example, autoregressive dynamics) and does not decouple spatial variation. Applying PS verbatim therefore underuses spatial correlation in large-scale fading and neglects informative environmental priors, which limits performance under joint spatiotemporal A2G variability.

To overcome these limitations, recent works incorporate physical prior information about A2G channels into online learning. A representative paradigm is the channel knowledge map (CKM), which encodes environmental knowledge for communication systems [8], [9]. CKM-based approaches reduce signaling overhead and improve handover reliability by replacing frequent channel probing with environment-aware predictions. However, constructing and maintaining an accurate CKM requires extensive measurements and incurs high computational cost. Outdated maps may mislead decision-making and degrade performance. Moreover, CKM mainly captures large-scale propagation characteristics and cannot reflect small-scale dynamics such as Doppler effects and multipath fading.

To address these limitations, we develop a hierarchical framework that integrates radio-map priors with the PS architecture. The A2G channel is decoupled into large-scale spatial and small-scale temporal structures to capture environment-dependent variation. Within this framework, we propose environment-aware predictive sampling (EA-PS), which employs a dual-sampling mechanism to jointly track large-scale shadowing residuals and small-scale fading dynamics. The design supports robust channel tracking with computationally efficient posterior updates. The main contributions of this paper are summarized as follows:

- We develop an environment-aware learning framework for UAV communications by incorporating radio map priors into a hierarchical channel decomposition. This approach effectively accounts for both large-scale spatial

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62471311 and No. 62271325, and in part by the Guangdong Basic and Applied Basic Research Foundation under Project 2024B1515020089, and in part by Shenzhen Science and Technology Program under Project JCYJ20250604181912016. (Corresponding author: Rui Zhang.)

characteristics and small-scale temporal dynamics while maintaining high learning efficiency.

- We propose the EA-PS algorithm, which utilizes a dual-sampling mechanism to jointly track large-scale shadowing residuals and small-scale fading. The algorithm enables efficient posterior updates and demonstrates strong robustness in non-stationary A2G environments.
- We establish a Bayesian regret upper bound via a novel spatiotemporal information decomposition and validate the framework numerically. Simulations indicate that EA-PS attains roughly 23% higher cumulative reward and over 50% lower regret compared with existing non-stationary bandit baselines.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first establish the A2G channel model for UAV-GBS communications. Building on this model, we then describe the mapping of the dynamic base station switching problem into an MAB optimization task.

A. Hierarchical Channel Model

We consider a communication system comprising a UAV and multiple GBSs. Owing to the UAV's mobility and frequent topology changes, the A2G links exhibit pronounced time-varying, non-stationary behavior. Specifically, large-scale effects appear as rapid path loss fluctuations and shadowing dynamics, while small-scale effects are dominated by large Doppler shifts and the temporal evolution of multipath components. Because jointly modeling both scales remains formidable and compute-intensive, an accurate hierarchical mathematical description is essential for reliable A2G communications. In this work, we adopt a hierarchical channel model that decomposes the wireless channel into large-scale and small-scale components, expressed as

$$H_{a,k}(t) = \sqrt{\beta_a(t)} H_{a,k}^{\text{small}}(t), \quad (1)$$

where $H_{a,k}(t)$ denotes the aggregate channel gain for the k -th subcarrier associated with GBS a , $\beta_a(t)$ is the large-scale fading coefficient associated with GBS a , and $H_{a,k}^{\text{small}}(t)$ represents the small-scale fading component.

At the large scale, we leverage a radio map to provide prior knowledge of channel characteristics, including path loss and shadowing. The large-scale fading coefficient associated with GBS a is expressed as

$$\beta_a(t) = \zeta(\mathbf{x}_t) \cdot \chi(\mathbf{x}_t), \quad (2)$$

where $\zeta(\mathbf{x}_t)$ and $\chi(\mathbf{x}_t)$ characterize the path loss and shadowing effects as functions of the UAV position $\mathbf{x}_t$ at time t , respectively.

At the small scale, we focus on fast fading induced by multipath propagation. We assume that both the UAV and the GBSs are equipped with a single antenna. The system employs orthogonal frequency-division multiplexing (OFDM), where the total bandwidth is divided into K orthogonal subcarriers.

The small-scale fading for GBS a on subcarrier k at time t takes the multipath representation

$$H_{a,k}^{\text{small}}(t) = \sum_{p=1}^{N_p} h_{a,p}(t) e^{-j2\pi f_k \tau_p}, \quad (3)$$

where N_p denotes the number of multipath components, $h_{a,p}(t)$ is the complex channel gain of the p -th path at time step t , f_k is the frequency of the k -th subcarrier, and τ_p is the delay of the p -th path.

B. Multi-armed Bandit Model

To characterize channel quality, we adopt the received signal power as the performance indicator, aligning with classical received signal strength formulations [10]. Under the standard sequential protocol, at time t , the agent selects an arm A_t , and then observes reward R_{t+1,A_t} . For a generic arm a , we define reward as the average channel power gain over K subcarriers, given by

$$R_{t+1,a} = \frac{1}{K} \sum_{k=1}^K |H_{a,k}(t)|^2 + \varepsilon_{t+1,a} = \beta_a(t) G_{t+1,a} + \varepsilon_{t+1,a}, \quad (4)$$

where $H_{a,k}(t)$ denotes the complex channel gain of the k -th subcarrier, $\varepsilon_{t+1,a}$ denotes measurement noise, and $G_{t+1,a}$ is the small-scale power gain given by

$$G_{t+1,a} = \frac{1}{K} \sum_{k=1}^K |H_{a,k}^{\text{small}}(t)|^2. \quad (5)$$

Let $\mathcal{A}$ denote the finite set of available arms, where each $a \in \mathcal{A}$ identifies a candidate GBS, and let $A_t \in \mathcal{A}$ be the arm played at time t . Maximizing expected cumulative reward over horizon T is expressed by

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} R_{t+1,A_t} \right], \quad (6)$$

where π denotes an online decision policy that maps the history H_t to an action A_t .

Optimizing UAV communication performance via GBS selection amounts to maximizing the cumulative reward through sequential decision-making. This problem can be naturally formulated as an MAB. In this context, the available GBSs are mapped to the arms of the bandit, and the average channel power serves as the reward signal.

III. ENVIRONMENT-AWARE PREDICTIVE SAMPLING

Building on Section II, we detail the Gaussian process (GP) model for large-scale fading and the first-order auto-regressive (AR(1)) characterization of small-scale variation, summarize their posterior updates, and then present the EA-PS procedure.

A. Decoupled Modeling

Characterizing both scales with a unified temporal stochastic model is analytically challenging due to the fundamental spatiotemporal discrepancy. Given the disparity between spatial and temporal dynamics, we decouple the channel characterization into scale-specific mathematical treatments.

1) *Large-scale Modeling*: The large-scale fading β_a associated with the link between the UAV and GBS a is factorized according to the radio-map prior plus a log-domain residual,

$$\beta_a(t) = \beta_{\text{map},a}(t) \cdot \exp(\delta_a(t)), \quad (7)$$

where $\beta_{\text{map},a}(t)$ encodes radio-map predictable large-scale gain between the UAV and GBS a , while $\delta_a(t)$ denotes the residual map error attributable to mismatched shadowing. The spatially dependent shadowing residual $\delta_a(\mathbf{x})$ is modeled as a GP [11]. Let $\mathbf{x}_t \in \mathcal{X}$ denote the UAV location at time t so that $\delta_a(t) \triangleq \delta_a(\mathbf{x}_t)$, let $\mathcal{T}_t^a$ denote the index set associated with arm a at horizon t , and let $\mathbf{y}_t^a = [y_{a,\tau_1}, \dots, y_{a,\tau_{|\mathcal{T}_t^a|}}]^\top$ stack the noisy residual observations of that arm, where $y_{a,t} = \delta_a(\mathbf{x}_t) + \epsilon_{a,t}^y$ and $\epsilon_{a,t}^y \sim \mathcal{N}(0, \sigma_y^2)$. The GP regression inputs are summarized by the design set expressed as

$$\mathcal{D}_t^a = \{(\mathbf{x}_\tau, y_{a,\tau}) : \tau \leq t, a_\tau = a\}. \quad (8)$$

Given the dataset $\mathcal{D}_t^a$, the posterior distribution of the residual at a query location $\mathbf{x}_*$ follows a Gaussian distribution $\mathcal{N}(\mu_a^{GP}(\mathbf{x}_*), \alpha_a^2(\mathbf{x}_*))$, where the posterior mean and variance are given by

$$\mu_a^{GP}(\mathbf{x}_*) = \mathbf{k}_*^\top (\mathbf{K}_t^a + \epsilon_n \mathbf{I})^{-1} \mathbf{y}_t^a, \quad (9)$$

and

$$\alpha_a^2(\mathbf{x}_*) = \max(\epsilon_v, k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (\mathbf{K}_t^a + \epsilon_n \mathbf{I})^{-1} \mathbf{k}_*), \quad (10)$$

respectively. Here, $\epsilon_n = \sigma_y^2 + \epsilon_{\text{jit}}$ combines the residual-observation noise variance with a small numerical jitter, while ϵ_v is a variance floor. The vector $\mathbf{k}_*$ gathers cross covariances between $\mathbf{x}_*$ and training inputs, and $\mathbf{K}_t^a \in \mathbb{R}^{|\mathcal{T}_t^a| \times |\mathcal{T}_t^a|}$ denotes the covariance matrix with entries $[\mathbf{K}_t^a]_{i,j} = k(\mathbf{x}_{\tau_i}, \mathbf{x}_{\tau_j})$. The stationary kernel evaluated on $\mathbf{x}_i, \mathbf{x}_j$ reads

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha_0^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2d_{\text{cor}}^2}\right), \quad (11)$$

where d_{cor} is the correlation distance.

To mitigate the $O(t^3)$ scaling of full-batch GP inference, we employ a sliding window that retains only the W_{GP} most recent residuals so that $|\mathcal{T}_t^a| \leq W_{\text{GP}}$ and each update requires $O(W_{\text{GP}}^3)$ operations. This windowed construction preserves real-time viability while retaining accuracy. Across $|\mathcal{A}|$ arms, GP prediction costs $O(|\mathcal{A}|W_{\text{GP}}^2)$ per slot and cached kernel matrices require $O(|\mathcal{A}|W_{\text{GP}}^2)$ memory; the arm-wise Kalman recursions cost $O(|\mathcal{A}|)$.

2) *Small-scale Modeling*: In the proposed channel simulator, whenever K is large enough for the multipath aggregates to resemble Gaussian fluctuations, $G_{t+1,a}$ is approximated by a Gaussian surrogate with nominal mean $\theta_{t,a}$ and variance σ_a^2 .

$$G_{t+1,a} = \theta_{t,a} + Z_{t+1,a}, \quad (12)$$

where $Z_{t+1,a}$ represents an independent zero-mean noise term with variance σ_a^2 .

To capture temporal dynamics, the latent state $\vartheta_{t,a}$ (equivalent to $\theta_{t,a}$), which represents the mean small-scale power gain, is modeled as an AR(1) process:

$$\vartheta_{t+1,a} = (1 - \gamma_a)c_a + \gamma_a\vartheta_{t,a} + W_{t+1,a}, \quad (13)$$

where γ_a and c_a denote the damping factor and the steady-state mean level of the recursion, whereas $W_{t+1,a}$ denotes independent Gaussian noise with deterministic variance η_a^2 for $\eta_a \in \mathbb{R}^+$. Conditional modeling of instantaneous power gains adopts the shorthand law expressed as

$$G_{t+1,a} \sim \mathcal{N}(\theta_{t,a}, \sigma_a^2). \quad (14)$$

Accordingly, conditioned on H_t , the posterior for θ_t satisfies $\theta_t | H_t \sim \mathcal{N}(\mu_t, \Sigma_t)$, where μ_t and Σ_t can be derived analytically using a Kalman filter.

To instantiate PS-compatible predictive sampling despite drift in θ , we employ the adjusted Gaussian posterior $\mathcal{N}(\tilde{\mu}_t, \tilde{\Sigma}_t)$ advocated in [7]. The matrices Σ_t and $\tilde{\Sigma}_t$ are diagonal, entries $\sigma_{t,a}^2$ and $\tilde{\sigma}_{t,a}^2$ track arm-wise uncertainty, and $\tilde{\Sigma}_t = \text{diag}(\tilde{\sigma}_{t,1}^2, \dots, \tilde{\sigma}_{t,A}^2)$. The adjusted moments satisfy the map given by

$$\tilde{\mu}_t = \mu_t, \quad \text{and} \quad \tilde{\sigma}_{t,a}^2 = \frac{\gamma_a^2 \sigma_{t,a}^4}{\gamma_a^2 \sigma_{t,a}^2 + x_a^*}, \quad (15)$$

where

$$x_a^* = \frac{\eta_a^2 + \sigma_a^2 - \gamma_a^2 \sigma_a^2 + \sqrt{(\eta_a^2 + \sigma_a^2 - \gamma_a^2 \sigma_a^2)^2 + 4\gamma_a^2 \eta_a^2 \sigma_a^2}}{2} \quad (16)$$

Hence reversible sampling induces the latent-mean equivalences expressed by

$$\begin{aligned} P(\mathbb{E}[R_{t+1,A_t} | H_t, R_{t+2:\infty}] \in \cdot | H_t) \\ = P(\mathbb{E}[\theta_t | H_t, R_{t+2:\infty}] \in \cdot | H_t), \end{aligned} \quad (17)$$

where the history augmentation H_{t+1} is expressed as

$$\begin{aligned} H_{t+1} = (\mathbf{x}_0, A_0, \beta_{\text{map}}(\mathbf{x}_0, A_0), y_{A_0,0}, R_{1,A_0}, \\ \dots, \mathbf{x}_t, A_t, \beta_{\text{map}}(\mathbf{x}_t, A_t), y_{A_t,t}, R_{t+1,A_t}). \end{aligned} \quad (18)$$

B. EA-PS Algorithm

We perform independent posterior sampling for the large-scale residual component and the small-scale fading component, and combine them multiplicatively in the decision rule. This design enables coherent exploration across both spatial and temporal domains.

Under the proposed hierarchical model, each component is updated according to its corresponding observation model, and the resulting decision rule remains consistent with the PS framework.

Building on the foregoing models, Algorithm 1 lists the EA-PS updates adopted in simulations. Consequently, the proposed method incorporates environment awareness without compromising the underlying mechanism and theoretical structure of the PS framework, thereby achieving both theoretical tractability and practical robustness in non-stationary UAV-ground communication scenarios.

Remark 1. *The scalar reward alone cannot identify the two multiplicative components. We therefore use selected-link CSI-assisted feedback: standard slow-fast-fading separation*

Algorithm 1 Environment-Aware Predictive Sampling

```

1: for  $t = 0, 1, \dots, T - 1$  do
2:   Sample  $\hat{\theta}_t \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\Sigma}_t)$ 
3:   Sample  $\hat{\delta}_{t,a} \sim \mathcal{N}(\mu_{t-1}^{GP}(\mathbf{x}_t, a), \alpha_{t-1}^2(\mathbf{x}_t, a))$  for all  $a \in \mathcal{A}$ 
4:   Select  $A_t \in \arg \max_{a \in \mathcal{A}} \beta_{\text{map},a}(t) \exp(\hat{\delta}_{t,a}) \hat{\theta}_{t,a}$ 
5:   Observe  $R_{t+1,A_t}$  and a large-scale estimate  $\hat{\beta}_{A_t}(t)$  of the selected link
6:   Form  $y_{A_t,t} = \ln(\hat{\beta}_{A_t}(t)/\beta_{\text{map},A_t}(t))$  and  $\hat{G}_{t+1,A_t} = R_{t+1,A_t}/\hat{\beta}_{A_t}(t)$ 
7:   Update  $(\tilde{\mu}_{t+1}, \tilde{\Sigma}_{t+1})$  using  $\hat{G}_{t+1,A_t}$ , the Kalman filter, and (13)
8:   Update the GP posterior of arm  $A_t$  using  $(\mathbf{x}_t, y_{A_t,t})$  and (9)–(10)
9: end for

```

provides $\hat{\beta}_{A_t}(t) = \beta_{A_t}(t) \exp(\epsilon_{A_t,t}^y)$ for the selected GBS. Combining this estimate with (7) gives

$$y_{A_t,t} = \ln \frac{\hat{\beta}_{A_t}(t)}{\beta_{\text{map},A_t}(t)} = \delta_{A_t}(\mathbf{x}_t) + \epsilon_{A_t,t}^y. \quad (19)$$

Thus, $\delta_{A_t}(\mathbf{x}_t)$ is the systematic radio-map mismatch, whereas $\epsilon_{A_t,t}^y$ is the zero-mean observation error. The same decomposition yields $\hat{G}_{t+1,A_t} = R_{t+1,A_t}/\hat{\beta}_{A_t}(t)$ for the Kalman update, with decomposition and reward noise absorbed into its observation-noise term. Only the selected link is measured, so no information about unselected GBSs is revealed.

IV. PERFORMANCE ANALYSIS

In this section, we analyze the regret of the proposed EA-PS algorithm from a Bayesian perspective.

A. Regret Formulation

We use $R_{t+1,a}$ to denote the realized reward obtained by selecting action a at slot t . Combining (4) and (7) yields the instantaneous form expressed as $R_{t+1,a} = \beta_{\text{map},a}(t) \exp(\delta_{t,a}) \theta_{t,a} + \varepsilon_{t+1,a}$, where $\varepsilon_{t+1,a}$ collects thermal noise and measurement inaccuracies. Concatenating the arm-wise residuals and means into $\mathbf{s}_t = [\delta_t^\top, \theta_t^\top]^\top \in \mathcal{S}$, with $\delta_t = (\delta_{t,a})_{a \in \mathcal{A}}$ and $\theta_t = (\theta_{t,a})_{a \in \mathcal{A}}$, yields the instantaneous Bayes-optimal selection $A_t^* = \arg \max_{a \in \mathcal{A}} \mathbb{E}[R_{t+1,a} | H_t, \mathbf{s}_t]$.

Following Russo et al. [12], cumulative regret admits the canonical information-ratio bound. The instantaneous information ratio is expressed as

$$\Gamma_t = \frac{\mathbb{E}[R_{t+1,A_t^*} - R_{t+1,A_t} | H_t]^2}{I(\mathbf{s}_t; (A_t, R_{t+1,A_t}) | H_t)}, \quad (20)$$

The conditional sub-Gaussian requirement concerns the observation noise rather than the multiplicative latent mean. Conditioned on $\mathbf{s}_t$, $g(\mathbf{s}_t) = \beta_{\text{map}} \exp(\delta) \theta$ is deterministic and $R_{t+1,A_t} - \mathbb{E}[R_{t+1,A_t} | \mathbf{s}_t] = \varepsilon_{t+1,A_t}$, which is assumed zero-mean and conditionally σ^2 -sub-Gaussian. Compact truncation of δ and θ is used only to keep g bounded; discarded Gaussian tails contribute the stated negligible slack. For all $t \in \mathbb{Z}_+$, the

sampling prescription implies $\mathbb{P}(A_t^* \in \cdot | H_t) = \mathbb{P}(A_t \in \cdot | H_t)$ and $A_t^* \perp A_t | H_t$. Conditioning further on $\mathbf{s}_t$ yields $(A_t, R_{t+1,A_t}) \perp A_t^*$ so that $I(\mathbf{s}_t; (A_t, R_{t+1,A_t}) | H_t) \geq I(A_t^*; (A_t, R_{t+1,A_t}) | H_t)$. The following lemma records the resulting Kullback-Leibler inequality.

Lemma 1. Assume each $R_{t+1,a}$ is conditionally σ^2 -sub-Gaussian given $\mathbf{s}_t$. Then the posterior regret dispersion obeys

$$\begin{aligned} & \mathbb{E}[R_{t+1,A_t^*} - R_{t+1,A_t} | H_t]^2 \\ & \leq 2\sigma^2 |\mathcal{A}| \times I(\mathbf{s}_t; (A_t, R_{t+1,A_t}) | H_t). \end{aligned} \quad (21)$$

Applying the mutual-information chain rule to $\mathbf{s}_t = (\delta_t, \theta_t)$ yields the telescoping decomposition expressed by

$$\begin{aligned} I(\mathbf{s}_t; R_{t+1,A_t} | H_t) &= I(\delta_t; R_{t+1,A_t} | H_t) \\ & \quad + I(\theta_t; R_{t+1,A_t} | H_t, \delta_t), \end{aligned} \quad (22)$$

where δ_t and θ_t are the components of the latent state defined previously. This decomposition allows for a decoupled analysis of the information gain contributed by the spatial bias and the time-varying parameter.

B. Regret Analysis

In the following, we derive individual upper bounds for the information ratio and the cumulative mutual information, respectively, to establish the final cumulative regret bound.

1) *The Upper Bound of Information Ratio:* By employing the same truncation argument, δ_t can again be modeled as conditionally sub-Gaussian jointly with Gaussian θ_t and ε_{t+1,A_t} . Consequently, [12] supplies the analytic bound expressed as

$$\Gamma_t \leq 2|\mathcal{A}| \sigma^2, \quad (23)$$

where $|\mathcal{A}|$ denotes the cardinality of the action space.

2) *The Upper Bound of Cumulative Mutual Information:* Based on the aforementioned decomposition of mutual information, we derive the upper bound for the cumulative information gain by separately bounding the large-scale component, $I(\delta_t; R_{t+1,A_t} | H_t)$, and the small-scale component, $I(\theta_t; R_{t+1,A_t} | H_t, \delta_t)$. For the residual process δ , the cumulative mutual-information growth is bounded by the worst-case Gaussian maximizer Ψ_T associated with sampling T spatial locations, aligning with analyses of Gaussian-process upper-confidence sampling [13]. Specifically, we retain the slack expressed as

$$\sum_{t=1}^T I(\delta_t; R_{t+1,A_t} | H_t) \leq \max_{X_T \subseteq \mathcal{X}, |X_T|=T} I(\mathbf{y}_{X_T}; \mathbf{f}_{X_T}) = \Psi_T. \quad (24)$$

Fixing δ yields Gaussian θ , so Bayesian updates reduce to a linear Gaussian channel with capacity upper limits. The surrogate log-det bound reads

$$\sum_{t=1}^T I(R_{t+1,A_t}; \theta_t | H_t, \delta_t) \leq \frac{1}{2} \sum_{t=1}^T \log \left(1 + \frac{c_t^2 \sigma_{\text{pred},t}^2}{\sigma_\varepsilon^2} \right). \quad (25)$$

where $c_t \triangleq \beta_{\text{map},A_t}(t) \exp(\delta_{t,A_t})$ is the (time-varying) observation gain for the selected arm, $\sigma_\varepsilon^2 \triangleq \text{Var}(\varepsilon_{t+1,A_t})$, and

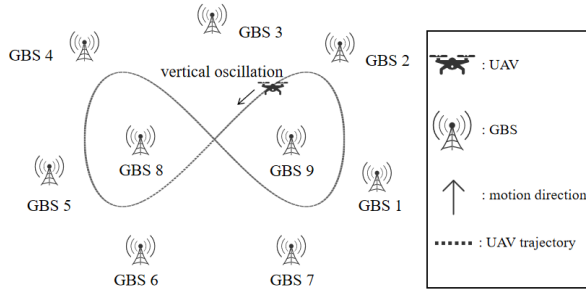


Fig. 1. Illustration of the simulation environment. A Lissajous trajectory is adopted for the UAV, with a time-varying altitude $z(t) = 100 + 30 \sin(0.02t)$.

$\sigma_{\text{pred},t}^2 = \gamma_{A_t}^2 \tilde{\sigma}_{t-1,A_t}^2 + \eta_{A_t}^2$ is the prediction variance under the AR(1) model. For brevity, define $\kappa_t \triangleq \frac{c_t^2 \sigma_{\text{pred},t}^2}{\sigma_\varepsilon^2}$.

Combining (23), (24), and (25) produces the pooled Bayesian regret slack expressed by

$$\mathbb{E}[\text{Regret}(T)] \leq \sqrt{2|\mathcal{A}|\sigma^2 T \left(\Psi_T + \frac{1}{2} \sum_{t=1}^T \log(1 + \kappa_t) \right)}. \quad (26)$$

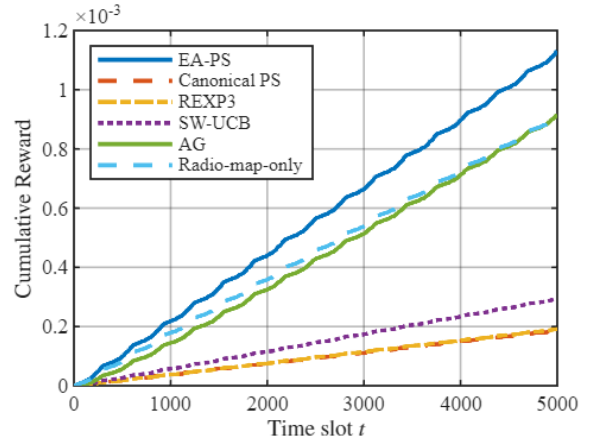
The result shows that the regret is jointly controlled by temporal uncertainty from small-scale fading and spatial mismatch of large-scale shadowing. This decomposition offers a coherent interpretation of simultaneous exploration across time and space.

V. NUMERICAL RESULTS

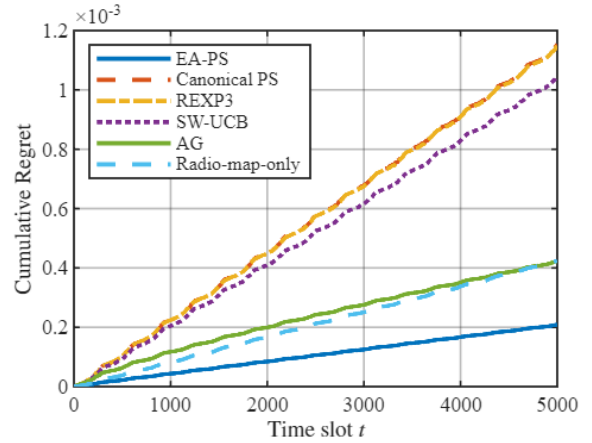
We emulate a QuaDRiGa deployment under 3GPP TR 38.901 Urban Macro assumptions. The waveform uses a 2.5 GHz carrier, orthogonal frequency-division multiplexing (OFDM) with 128 subcarriers spaced by 15 kHz, and measurement variances $\sigma_n^2 = 0.01$ alongside Kalman driving noise $\sigma_z^2 = 0.05$. Path-loss tables follow the standard QuaDRiGa map, while correlated log-normal shadowing injects 8 dB fluctuations with 50 m decorrelation footprints.

Remark 2. *The 3GPP UMa deployment and Lissajous trajectory jointly create rapidly varying peripheral links and relatively stable central links, thereby exposing the exploration–stability tradeoff targeted by EA-PS. Other deployments can be accommodated by replacing the radio map without changing the decision rule; however, the reported numerical gains are specific to the present configuration, and broader validation under map aging and diverse trajectories is left for future work.*

As illustrated in Fig. 1, the simulation area contains nine GBSs, corresponding to Arms 1 through 9. Arms 1 to 7 are distributed along the periphery of the UAV’s Lissajous trajectory, exhibiting strong non-stationarity. The remaining two GBSs (Arms 8 and 9) are located near the center of the area, serving as relatively stable options. This topological design allows us to evaluate the tradeoff between selecting high-



(a) Cumulative reward $\uparrow$



(b) Cumulative regret $\downarrow$

Fig. 2. Performance comparison of different algorithms.

reward but highly non-stationary arms and selecting stable but potentially suboptimal arms.

We benchmark EA-PS against canonical PS [7], sliding-window upper-confidence-bound (SW-UCB) bandits [14], adaptive greedy (AG), REXP3 [15], plus a heuristic that freezes decisions to radio-map forecasts only in Fig. 2. Across $T = 5000$ slots, EA-PS secures roughly 23% higher cumulative reward than the strongest baseline (AG) while reducing cumulative regret by more than 50%. In the ablation study, EA-PS outperforms both the canonical PS and radio-map-only baselines. Specifically, its cumulative reward is approximately 24% higher than that of the radio-map-only scheme and more than four times that of canonical PS. Furthermore, EA-PS achieves over 50% and 80% reductions in cumulative regret compared to the radio-map-only and canonical PS baselines, respectively. These performance limitations in the baselines stem from their neglect of either large-scale or small-scale fading effects. In contrast, by leveraging an environment-aware learning framework, EA-PS effectively captures both channel characteristics.

Table I reports the empirical arm-selection frequencies. Canonical PS collapses onto Arm 1 because its purely tem-

TABLE I
ARM SELECTION PROBABILITIES OF EACH ALGORITHM

Algorithm	Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	Arm 6	Arm 7	Arm 8	Arm 9
EA-PS	0.0973	0.1052	0.0997	0.1033	0.1147	0.1071	0.1090	0.1316	0.1322
Canonical PS	0.9984	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
REXP3	0.1107	0.1113	0.1111	0.1113	0.1109	0.1114	0.1108	0.1113	0.1113
SW-UCB	0.1085	0.1096	0.1088	0.1090	0.1108	0.1106	0.1097	0.1162	0.1168
AG	0.1166	0.1073	0.0867	0.1115	0.1119	0.1103	0.1264	0.1253	0.1041
Radio-map-only	0.0962	0.0974	0.0946	0.1017	0.1138	0.1085	0.1039	0.1418	0.1419

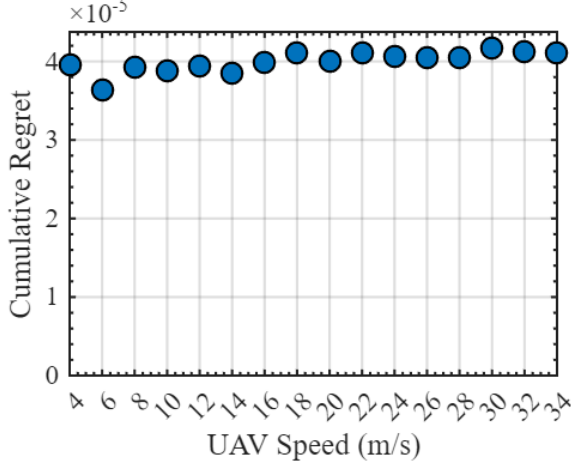


Fig. 3. Cumulative regret performance under various velocity configurations at $T = 2000$ ($\downarrow$).

poral surrogate cannot represent joint large- and small-scale structure. REXP3 and SW-UCB remain nearly uniform, over-exploring without recognizing the stability pockets around Arms 8 and 9. AG mitigates some regret by reducing exploration frequency, but it does not fundamentally address the underlying non-stationarity. In contrast, EA-PS strikes a favorable balance between reward maximization and selection stability. It effectively identifies high-reward arms while actively prioritizing the relatively stable Arms 8 and 9 when the optimal arms exhibit severe non-stationarity. By explicitly accounting for information persistence, EA-PS fundamentally adapts to the environmental dynamics, resulting in superior performance compared to the baseline algorithms.

To evaluate the robustness of EA-PS, we illustrate the cumulative regret of EA-PS at $T = 2000$ over multiple simulation scenarios versus UAV velocities in Fig. 3. It shows that the regret remains relatively stable across a wide velocity span, ranging from 4 m/s to 34 m/s, which validates the algorithm's robustness to the mobility-induced dynamics. This velocity sweep does not isolate AR(1) parameter mismatch; a controlled study that perturbs γ_a and η_a is left for future work.

VI. CONCLUSIONS

This work studied online GBS selection for UAVs over non-stationary A2G channels where rewards co-evolve with position-driven large-scale effects and mobility-induced small-scale fading. We introduced EA-PS, which couples radio-map priors with a GP residual for spatial shadowing mismatch and a Kalman-tracked AR(1) small-scale component, realized

through dual predictive sampling aligned with multiplicative rewards. The analysis yields a Bayesian regret bound in which cumulative information gain cleanly splits into GP spatial uncertainty and AR(1) temporal uncertainty. QuaDRiGa-based evaluations against canonical PS, SW-UCB, AG, REXP3, and a radio-map-only policy confirm substantial gains and explainable selection behavior: EA-PS tracks time-varying high-value links without collapsing exploration like PS, yet avoids uniformly wasteful probing of adaptive bandits. Future work will examine radio-map errors and aging, GP-window and AR(1) parameter sensitivity, diverse trajectories, and reward models incorporating SINR, achievable rate, and handover cost.

REFERENCES

- [1] E. Vinogradov, A. Saboor, Z. Cui, and A. Fakhreddine, "Spatially consistent air-to-ground channel modeling with probabilistic LOS/NLOS segmentation," in *Proc. IEEE 101st Veh. Technol. Conf. (VTC2025-Spring)*, Jun 2025, pp. 1–5.
- [2] Y. Liu, C.-X. Wang, H. Chang, Y. He, and J. Bian, "A novel non-stationary 6G UAV channel model for maritime communications," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 2992–3005, 2021.
- [3] M. Banagar and H. S. Dhillon, "Fundamentals of wobbling and hardware impairments-aware air-to-ground channel model," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 12, pp. 17946–17962, 2024.
- [4] E. Cavenaghi, G. Sottocornola, F. Stella, and M. Zanker, "Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm," *Entropy*, vol. 23, no. 3, p. 380, Mar 2021.
- [5] M. A. Abdul Careem and A. Dutta, "Real-time prediction of non-stationary wireless channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7836–7850, 2020.
- [6] M. B. Booth, V. Suresh, N. Michelusi, and D. J. Love, "Multi-armed bandit beam alignment and tracking for mobile millimeter wave communications," *IEEE Communications Letters*, vol. 23, no. 7, pp. 1244–1248, Jul 2019.
- [7] Y. Liu, B. Van Roy, and K. Xu, "Nonstationary bandit learning via predictive sampling," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 206. PMLR, Apr 2023, pp. 6215–6244.
- [8] Y. Zeng, J. Chen, J. Xu, D. Wu, X. Xu, S. Jin, X. Gao, D. Gesbert, S. Cui, and R. Zhang, "A tutorial on environment-aware communications via channel knowledge map for 6G," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 3, pp. 1478–1519, 2024.
- [9] Y. Qiu, X. Chen, K. Mao, X. Ye, H. Li, F. Ali, Y. Huang, and Q. Zhu, "Channel knowledge map construction based on a UAV-assisted channel measurement system," *Drones*, vol. 8, no. 5, p. 191, May 2024.
- [10] I. Santos, J. Murillo-Fuentes, and P. M. Djurić, "Recursive estimation of dynamic RSS fields based on crowdsourcing and Gaussian processes," *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1152–1162, 2019.
- [11] R. Miyagusuku, A. Yamashita, and H. Asama, "Precise and accurate wireless signal strength mappings using Gaussian processes and path loss models," *Robotics and Autonomous Systems*, vol. 103, pp. 134–150, 2018.
- [12] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling," *Journal of Machine Learning Research*, vol. 17, no. 68, pp. 1–30, 2016.
- [13] D. Janz, D. Burt, and J. Gonzalez, "Bandit optimisation of functions in the Matérn kernel RKHS," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 26–28 Aug 2020, pp. 2486–2495.
- [14] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Algorithmic Learning Theory*, J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 174–188.
- [15] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration-exploitation in a multi-armed bandit problem with non-stationary rewards," *Stochastic Systems*, vol. 9, no. 4, pp. 319–337, 2019.